ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Ανάκτηση Boole

Ακαδημαϊκό Έτος 2022-2023

Περιεχόμενο

1. Μερικές βασικές έννοιες

2. Ένα απλό ΣΑΠ

   Μια μικρή εισαγωγή στο απλούστερο μοντέλο αναζήτησης (Boolean) (Κεφάλαιο 1 του βιβλίου)

   Ένα απλό σύστημα ΑΠ (βασικές δομές δεδομένων και παραδείγματα ερωτημάτων)

Boolean Ανάκτηση

Τι είναι η Ανάκτηση Πληροφορίας (Information Retrieval);

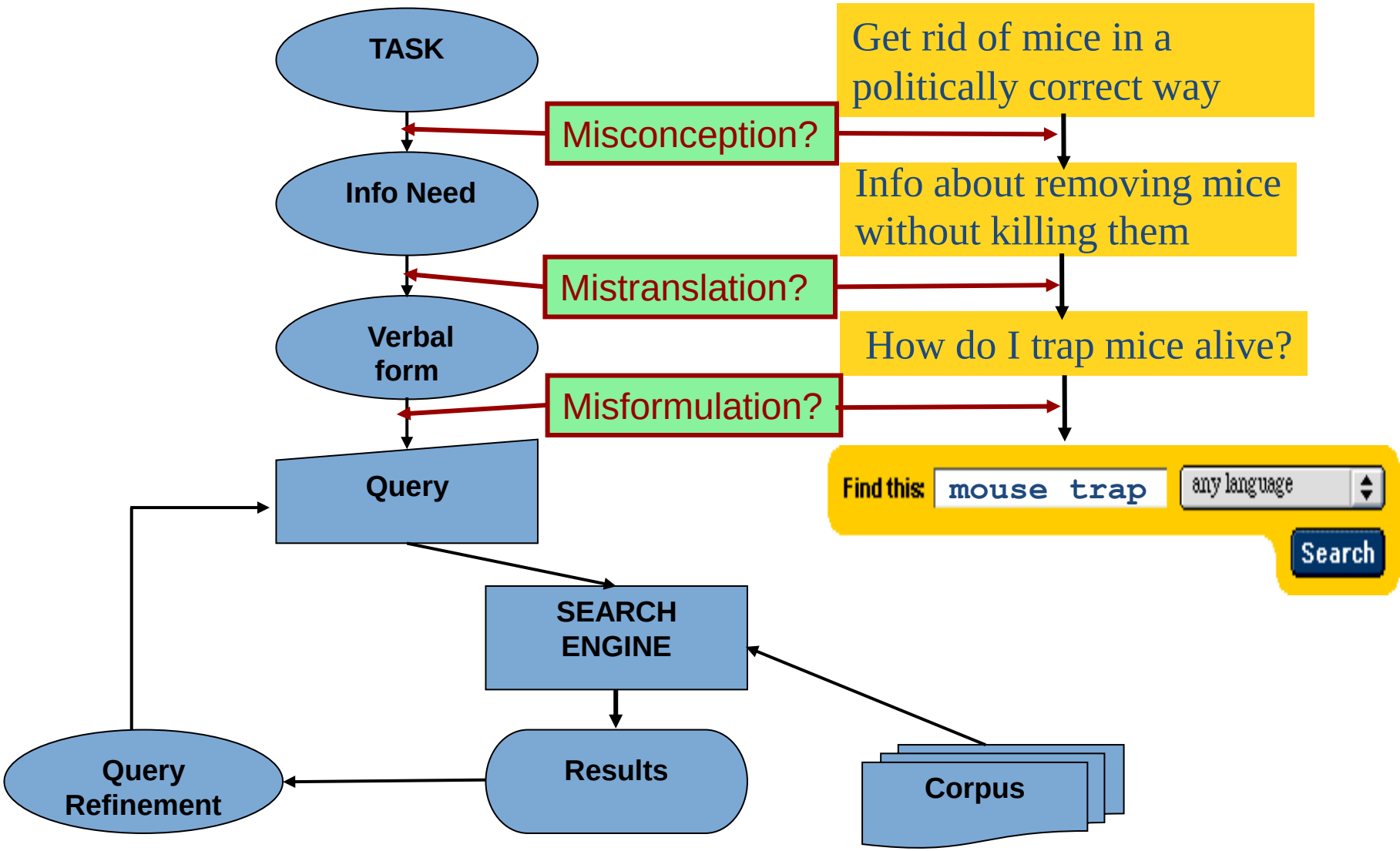
Βασικές Έννοιες

Συλλογή (Collection - corpus): Σύνολο από έγγραφα

Στόχος: Ανάκτηση των εγγράφων που περιέχουν πληροφορία που είναι συναφής (relevant) με την ανάγκη πληροφόρησης (information need) του χρήστη και τον βοηθά να ολοκληρώσει κάποιο έργο (task)

- ✓ Διαφορά μεταξύ: information need και ερωτήματος (query)
- ✓ Ad hoc retrieval

Το κλασικό μοντέλο αναζήτησης (search model)



Βασικές έννοιες

Αποτέλεσμα σε διάταξη με βάση τη συνάφεια

Αξιολόγηση:

- πέρα από την απόδοση (efficiency) αποτελεσματικότητα (effectiveness)

Αποτελεσματικότητα (effectiveness): Πόσο *καλά (χρήσιμα)* είναι τα έγγραφα που ανακτήθηκαν;

Αποτελεσματικότητα

- *Ακρίβεια (Precision)*: Το ποσοστό των εγγράφων που ανακτήθηκαν που είναι συναφή με την ανάγκη πληροφόρησης του χρήστη
- *Ανάκληση (Recall)* : Το ποσοστό των συναφών με την ανάγκη πληροφόρησης του χρήστη εγγράφων της συλλογής που ανακτήθηκαν από το σύστημα
  - Περισσότερα στο μέλλον

παράδειγμα Συλλογή 1000 έγγραφα
υπάρχουν 20 συναφή έγγραφα για το ερώτημα
το ΣΑΠ μας επιστρέφει 30 έγγραφα { 17 συναφή
13 μη συναφή

Βασικά Βήματα

(προεπεξεργασία)

- Σύλλεξε τα έγγραφα
- Ανάλυση
- Κατασκεύασε βοηθητικές δομές – ευρετήρια

(λειτουργία)

- Επεξεργασία ερωτήσεων

Αρχικά θα δούμε την απλούστερη μορφή:

Boolean retrieval

Boolean μοντέλο

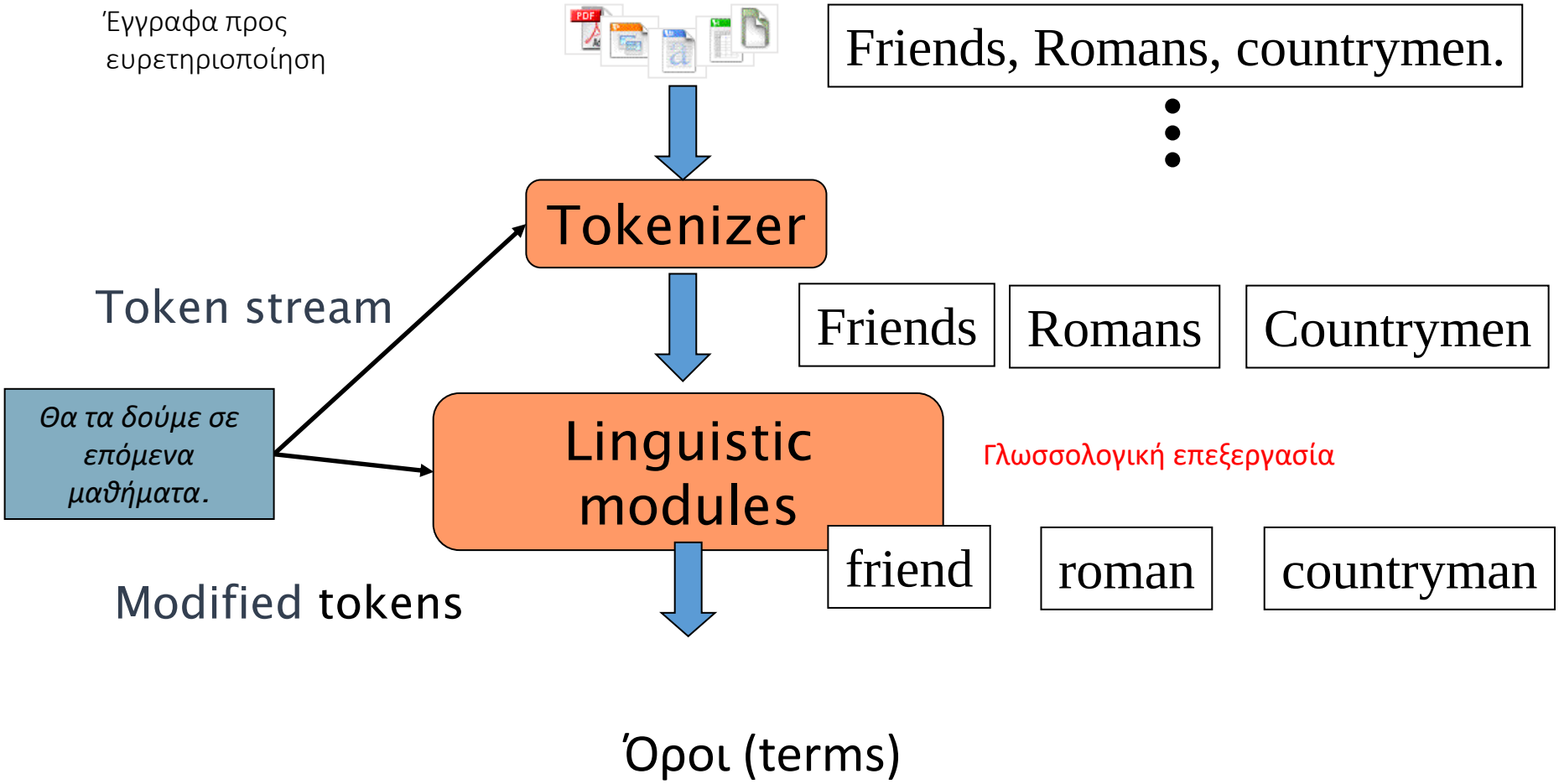
- Οι χρήστες διατυπώνουν ερωτήματα με τη μορφή *Boolean εκφράσεων*, δηλαδή όρων συνδυασμένων με AND, OR και NOT

- Επιστρέφονται ως απάντηση όλα τα κείμενα που ικανοποιούν το ερώτημα *χωρίς διάταξη*
  - *Δυαδική συνάφεια (συναφές, μη συναφές)*

- Κείμενο ως σύνολο όρων

Υπόθεση: έχει γίνει ανάλυση και κάθε κείμενο είναι μια συλλογή από όρους

Ανάλυση


Έγγραφα προς ευρετηριοποίηση
Friends, Romans, countrymen.
Tokenizer
Token stream
Friends
Romans
Countrymen
Θα τα δούμε σε επόμενα μαθήματα.
Linguistic modules
Γλωσσολογική επεξεργασία
friend
roman
countryman
Modified tokens
Όροι (terms)

Αδόμητα δεδομένα το 1680

Shakespeare's Collected Works

Αδόμητα δεδομένα το 1680

- Ποια θεατρικά έργα του Shakespeare περιέχουν τις λέξεις ***Brutus*** *και* ***Caesar*** αλλά όχι τη λέξη ***Calpurnia***
  - Ερώτημα*: **Brutus** AND **Caesar** AND NOT **Calpurnia***
- Να διαβάσουμε όλα τα έργα σειριακά από την αρχή σημειώνοντας …
- Θα μπορούσαμε να κάνουμε `grep` σε όλα τα έργα για ***Brutus*** και ***Caesar,*** και να σβήσουμε τις γραμμές που περιέχουν τη λέξη ***Calpurnia***

Αδόμητα δεδομένα το 1680

- Γιατί όχι grep?
  - Αργό (για μεγάλες συλλογές)
  - Grep line-oriented, η ανάκτηση πληροφορίας document-oriented
  - *NOT* ***Calpurnia*** δεν είναι εύκολο
  - Επιπρόσθετη λειτουργικότητα (π.χ., βρες τη λέξη ***Romans*** κοντά στο ***countrymen***)
  - Διάταξη! Ranked retrieval (τα «καλύτερα» έγγραφα ανάμεσα σε αυτά που ικανοποιούν την ερώτηση)
    - Σε επόμενα μαθήματα ….

Θα προ-επεξεργαστούμε τα έγγραφα και θα δημιουργήσουμε ευρετήρια

Για να δούμε τα βασικά …

Boolean μοντέλο

Δυαδική μήτρα (πίνακας) σύμπτωσης M

Γραμμές: Term (όροι, λέξεις)

Στήλες: Document (έγγραφα, έργα)

$M[i, j] = 1$, αν ο όρος i εμφανίζεται στο έγγραφο j

0, αλλιώς

Term-document incidence matrix (μήτρα σύμπτωσης)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 αν το έργο περιέχει τη λέξη, 0 αλλιώς

Brutus *AND* ***Caesar*** *BUT NOT* ***Calpurnia***

Οι όροι και τα έγγραφα ως διανύσματα

Έχουμε ένα *δυαδικό διάνυσμα* για κάθε *όρο* και κάθε *έγγραφο*

- Για να απαντήσουμε στην ερώτηση: παίρνουμε τα διανύσματα για το ***Brutus, Caesar*** και το *συμπλήρωμα* του διανύσματος για το ***Calpurnia*** ➔ bitwise *AND*.

  110100 *AND* 110111 *AND* 101111 = 100100.

Οι απαντήσεις:

- Antony and Cleopatra, Act III, Scene ii

Agrippa [Aside to DOMITIUS ENOBARBUS]: Why, Enobarbus,
When Antony found Julius ***Caesar*** dead,
He cried almost to roaring; and he wept
When at Philippi he found ***Brutus*** slain.

- Hamlet, Act III, Scene ii

Lord Polonius: I did enact Julius ***Caesar*** I was killed i' the
Capitol; ***Brutus*** killed me.

παράδειγμα

d_1 a b c a

d_2 d a b

d_3 b a f

d_4 a e

d_5 d f

Μεγαλύτερες συλλογές

- Ας θεωρήσουμε N = 1 εκατομμύρια έγγραφα, το καθένα έχει περίπου 1000 (διακριτούς) όρους (~2-3 σελίδες βιβλίου).
- Έστω ότι ανάμεσα τους υπάρχουν M = 500K διακριτοί (*distinct)* όροι.

Πόσα κελιά έχει ο πίνακας σύμπτωσης;

A. 1 δισεκατομμύριο
B. 500 δισεκατομμύρια
C. 500 εκατομμύρια
D. 5 δισεκατομμύρια

Πόσα μη μηδενικά κελιά (δηλαδή 1) έχει ο πίνακας σύμπτωσης;

A. 1 δισεκατομμύριο
B. 1 εκατομμύριο
C. 50 δισεκατομμύρια
D. 50 εκατομμύρια

Ποσοστό των 1:

Πόσο είναι το μέγεθος του πίνακα;

- Ο 500K x 1M πίνακας έχει μισό τρισεκατομμύριο 0's και 1.
- Αλλά δεν έχει περισσότερα από ένα δισεκατομμύριο 1.
  - Ο πίνακας είναι εξαιρετικά αραιός (sparse) – τουλάχιστον το 99.8% είναι 0.
- Ποια είναι μια καλύτερη αναπαράσταση;

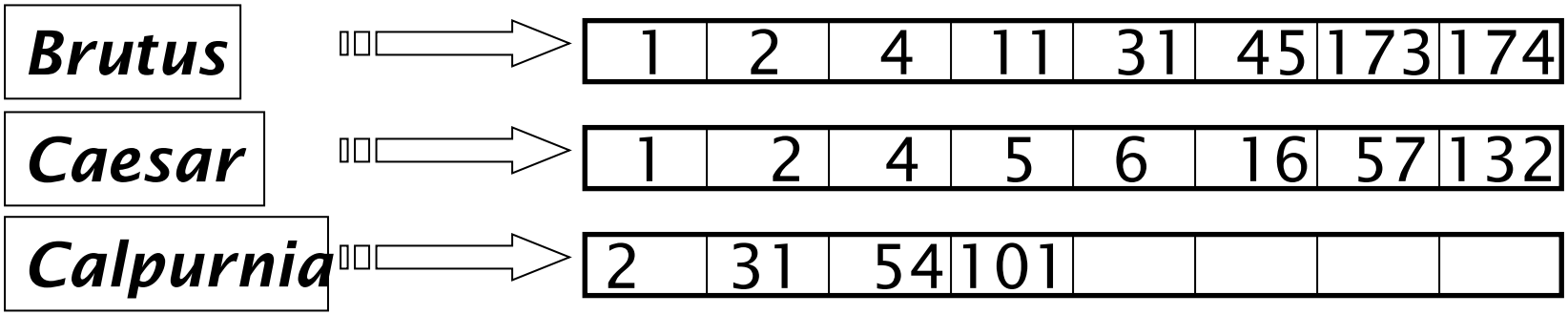
Αντεστραμμένο ευρετήριο ή αρχείο (Inverted index/file)

Για κάθε *όρο (term) t*, διατηρούμε μια λίστα με όλα τα έγγραφα που περιέχουν τον όρο.

- Κάθε έγγραφο χαρακτηρίζεται από ένα αναγνωριστικό εγγράφου (**docID**), πχ αριθμό που ανατίθεται σειριακά στα έγγραφα κατά τη δημιουργία τους

Αντεστραμμένο ευρετήριο

- Μπορούμε να χρησιμοποιήσουμε σταθερού μεγέθους arrays για αυτό?



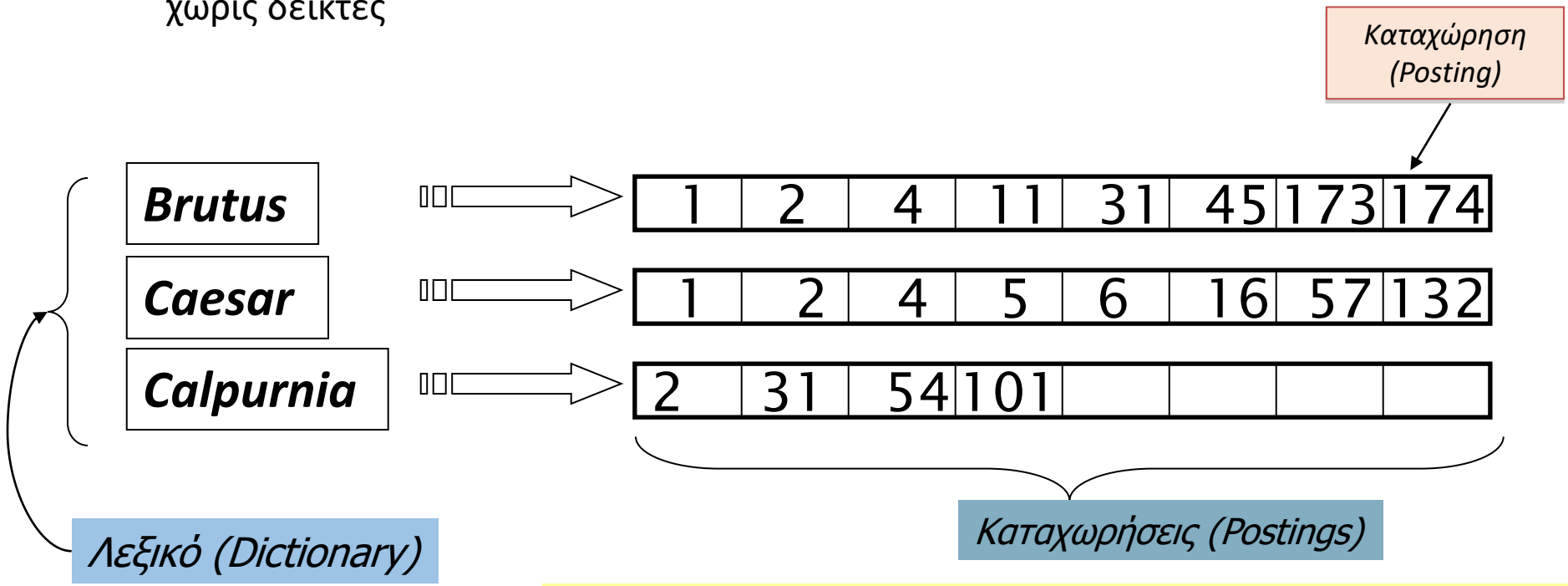
Τι γίνεται αν η λέξη ***Caesar*** προστεθεί στο έγγραφο 14?

Αντεστραμμένο ευρετήριο

- Χρειαζόμαστε μεταβλητού μεγέθους **λίστες καταχωρήσεων (postings lists)**

  Ποια δομή δεδομένων είναι κατάλληλη;

  - Στη μνήμη, απλά-διασυνδεδεμένες λίστες (skip lists) ή πίνακες μεταβλητού μήκους
  - Στο δίσκο, ως (συμπιεσμένες) συνεχόμενες ακολουθίες καταχωρήσεων χωρίς δείκτες



Σε διάταξη με βάση το docID (θα δούμε σε λίγο γιατί!).

παράδειγμα

d_1 a b c a

d_2 d a b

d_3 b a f

d_4 a e

d_5 d f

Βασική Ορολογία

- Αντεστραμμένο ευρετήριο (Inverted index)
- Λίστες καταχωρήσεων (posting lists) – μία για κάθε όρο
  - Καταχώρηση – ένα στοιχείο της λίστας
  - ✓ Κάθε λίστα είναι διατεταγμένη με το DocID
- Λεξιλόγιο (Vocabulary): το σύνολο των όρων
- Λεξικό (Dictionary) δομή δεδομένων για τους όρους
  - ✓ Αρχικά ας θεωρήσουμε αλφαβητική διάταξη

Το δημιουργούμε από πριν, θα δούμε πως

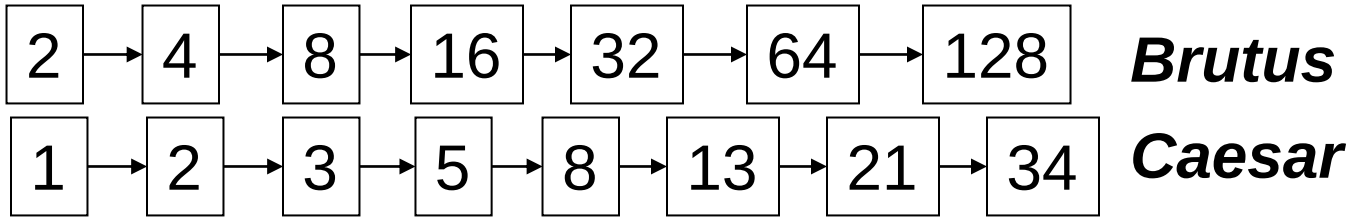
Φτιάξαμε το ευρετήριο, τώρα;

- Πως επεξεργαζόμαστε μια ερώτηση;
  - Αργότερα – τι άλλου είδους ερωτήσεις

Επεξεργασία ερωτήσεων: AND

Έστω η ερώτηση: ***Brutus*** *AND* ***Caesar***

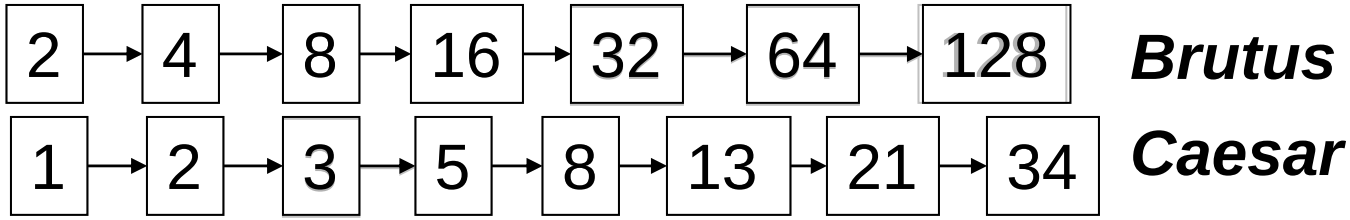
- Βρες το ***Brutus*** στο Λεξικό
  - Ανέκτησε τις καταχωρήσεις.
- Βρες το ***Caesar*** στο Λεξικό
  - Ανέκτησε τις καταχωρήσεις.
- Υπολογισμό της τομής του – ΠΩΣ;



Συγχώνευση (merge)

Κεφ. 1.3

- Διέσχισε τις δύο λίστες ταυτόχρονα, σε χρόνο γραμμικό (linear) στο συνολικό αριθμό των καταχωρήσεων



Αν τα μήκη των λιστών είναι *x* και *y*, η συγχώνευση παίρνει O(*x*+*y*) λειτουργίες.

Σημαντικό: οι καταχωρήσεις πρέπει να είναι διατεταγμένες με βάση το docID.

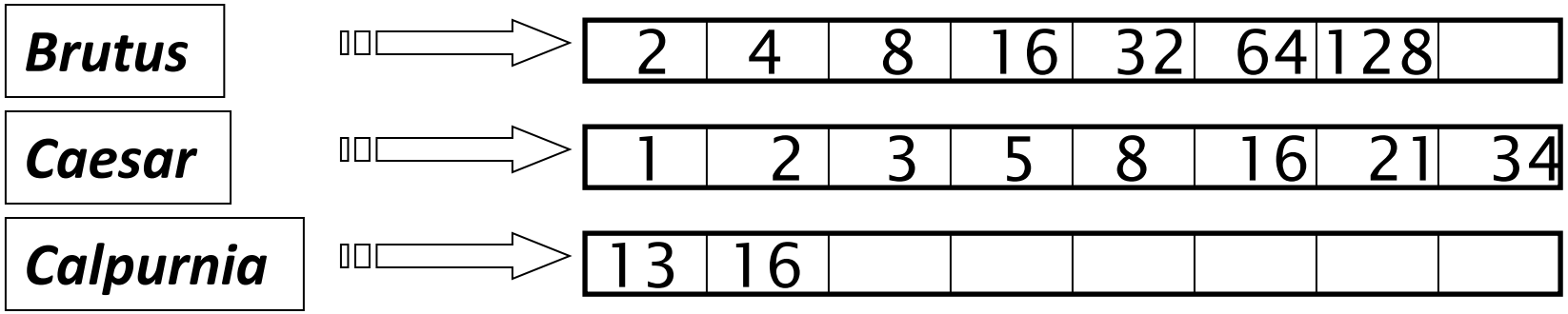
Ο αλγόριθμος συγχώνευσης

```
INTERSECT(p1, p2)
  answer ← ⟨ ⟩
  while p1 ≠ NIL and p2 ≠ NIL
  do if docID(p1) = docID(p2)
        then ADD(answer, docID(p1))
             p1 ← next(p1)
             p2 ← next(p2)
        else if docID(p1) < docID(p2)
                then p1 ← next(p1)
                else p2 ← next(p2)
  return answer
```

Βελτιστοποίηση ερωτήματος

- Ποια είναι *βέλτιστη σειρά* για την επεξεργασία ενός ερωτήματος;

- Έστω μια ερώτηση που είναι το *AND* *n* όρων.
- Για καθέναν από τους *n* όρους, βρες τις καταχωρήσεις του και εκτέλεσε το *AND* σε όλες.



Query: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

Βελτιστοποίηση ερωτήματος

- Επεξεργασία με αύξουσα συχνότητα:
  - *Ξεκίνησε με το μικρότερο σύνολο και συνέχισε μειώνοντας και άλλο το αποτέλεσμα*

Χρήση της συχνότητας εγγράφου στο λεξικό

Brutus	2	4	8	16	32	64	128	
Caesar	1	2	3	5	8	16	21	34
Calpurnia	13	16						

Εκτέλεση του ερωτήματος ως (***Calpurnia*** *AND* ***Brutus)*** *AND* ***Caesar***.

Βελτιστοποίηση ερωτήματος

Π.χ., *(**madding** OR **crowd**) AND (**ignoble** OR **strife**)*

- Βρες τη συχνότητα εγγράφου για όλους τους όρους.
- Εκτίμησε το μέγεθος κάθε *OR* (συντηρητικά: ως το άθροισμα των συχνοτήτων εγγράφου).
- Επεξεργασία του ερωτήματος κατά αύξουσα σειρά κάθε όρου.

Βελτιστοποίηση ερωτήματος

((A and B) and C) and D

- Κρατάμε *το ενδιάμεσο αποτέλεσμα στη μνήμη* και διαβάζουμε τη άλλη λίστα από το δίσκο
- Αρχικά, ενδιάμεσο αποτέλεσμα = A

Όταν πολλοί μεγάλες λίστες, εναλλακτικές για τον υπολογισμό τομής

- χρησιμοποιώντας δυαδική αναζήτηση στη μεγάλη λίστα (λογαριθμικός χρόνος)
- αποθήκευση μεγάλης λίστας ως hashtable (σταθερά)

- Δοκιμάστε το http://www.rhymezone.com/shakespeare/

Boolean ερωτήματα: Ακριβές ταίριασμα (Exact match)

- Το Boolean μοντέλο ανάκτησης απαντά ερωτήματα που είναι Boolean εκφράσεις:
  - Χρήση *AND, OR* και *NOT* για το συνδυασμό όρων
    - Θεωρούν κάθε έγγραφο ως ένα **σύνολο** όρων
    - Η αναζήτηση είναι ακριβής (precise): *ένα έγγραφο είτε ικανοποιεί τη συνθήκη είτε όχι.*
  - Ίσως, το απλούστερο μοντέλο
- Το βασικό μοντέλο σε εμπορικά συστήματα για 3 δεκαετίες (πριν τον web).
- Πολλά συστήματα ακόμα Boolean:
  - Email, library catalog, Mac OS X Spotlight

Η Google χρησιμοποιεί το Boolean μοντέλο ?

Τι είδαμε σήμερα

1- Βασικές έννοιες

2- Είδαμε ένα απλό σύστημα ανάκτησης πληροφορίας βασισμένο στο Boolean μοντέλο

α- ανάλυση εγγράφου
β- κατασκευάζουμε το ανεστραμμένο ευρετήριο
γ- το χρησιμοποιούμε για να απαντάμε σε ερωτήματα

ΤΕΛΟΣ 1ου Κεφαλαίου

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό των:

- ✓ *Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓ *Απόστολου Ν. Παπαδόπουλου , Ανάκτηση Πληροφορίας (Τμήμα Πληροφορικής, Αριστοτέλειο Πανεπιστήμιο)*

Extra slides

Άλλα ερωτήματα

Παράδειγμα: WestLaw

https://legal.thomsonreuters.com/en/products/westlaw

- Μεγάλο εμπορικό (συνδρομές επί πληρωμή) σύστημα
- Αναζήτηση σε νομικά κείμενα (άρχισε το1975, η διάταξη προστέθηκε το 1992)
- Boolean ερωτήματα και ερωτήματα φυσικής γλώσσας
- 40,000 databases for case laws, 60 countries (source: Wikipedia)

Παράδειγμα: WestLaw

- Παράδειγμα (απόσταση ανάμεσα στους όρους):
  - *Ανάγκη πληροφόρησης:* What is the statute of limitations in cases involving the federal tort claims act?
  - *Ερώτημα*:
    LIMIT! /3 STATUTE ACTION /S FEDERAL /2 TORT /3 CLAIM
    - /3 = within 3 words, /S = in same sentence

- Παράδειγμα (ανεκτική ανάκτηση):
  - *Ανάγκη πληροφόρησης:* Information on the legal theories involved in preventing the disclosure of trade secrets by employees formerly employed by a competing company
  - *Ερώτημα:*
    "trade secret" /s disclos! /s prevent /s employe!

Παράδειγμα: WestLaw

- Ακόμα ένα παράδειγμα:
  - Requirements for disabled people to be able to access a workplace
  - disabl! /p access! /s work-site work-place (employment /3 place)
- SPACE σημαίνει διάζευξη (disjunction)
- *Μακροσκελή, επακριβή ερωτήματα, τελεστές εγγύτητας (proximity operators), διατυπωμένα σταδιακά (διαφορά από web search)*
- *Boolean αναζήτηση χρησιμοποιείται ακόμα από πολλούς επαγγελματίες*
  - Ξέρεις ακριβώς τι παίρνεις ως απάντηση
- Αυτό δε σημαίνει ότι δουλεύει καλύτερα

Evidence accumulation

- 1 vs. 0 εμφάνιση ενός όρου αναζήτησης
  - 2 vs. 1 εμφανίσεις
  - 3 vs. 2 εμφανίσεις, κλπ.
  - Συχνά φαίνεται καλύτερο
- Χρειαζόμαστε και τη συχνότητα εμφάνισης του όρου στα έγγραφα

Ποιο περίπλοκη ημιδομημένη αναζήτηση

- *Title* is about Object Oriented Programming AND *Author* something like stro*rup
  - όπου * είναι ο wild-card τελεστής
- Θέματα:
  - Πως αντιμετωπίζουμε το “about”?
  - Πως γίνεται η κατάταξη?